Classification in Parkinson's disease
Parkinson's Disease

- The 2nd most common neurodegenerative disorder
- Impairs motor skills, speech, smell, cognition
- 1-3 sick per 1000
- >1% in individuals aged above 70
- Affects ~7M persons globally
- ~20 known mutations cover ~10% of cases
- >80% of the cases: idiopathic (IPD)
- Main cause: loss of neurons in the brain
- The last event in a chain of reactions
- Disease starts ~15 years before symptoms appear
- Still no cure, but therapies slow down progression
- Early diagnosis is important
Molecular markers of early Parkinson disease based on gene expression in blood

C. Scherzer et al, PNAS 2007

Lab for Neurogenomics, Harvard Med School
Setup

• Extracted whole blood RNA from 50 PD patients (at early disease stages) and 55 age-matched controls - including several disease controls = patients with other neurological diseases.
• Used Affy chips with 22,283 genes
• Training set: 66 samples: 31 patients, 17 healthy, 18 disease controls
Preprocessing

- Removed genes with low signal intensity (no value above the mean in at least one sample)
- Left 13,968 genes
- Ranked genes by absolute value of Pearson correlation with PD label
Feature selection procedure (1)

- Put one sample aside
  - Recompute correlation, rerank
  - For top n ranking genes do
  - Compute their average expression on PDs, controls - get 2 n-long vectors: PD template, control template
  - Use templates to evaluate the risk score of the left-out sample
- Predictive ability of the n genes: rank sum of the scores
- Choose n* maximizing the rank sum score over all samples

8 genes selected as the PD risk marker
Scores

- $s = \text{the LO sample}$
- **Risk score** $(s) = \text{Cor}(s, \text{PD template}) - \text{cor}(s, \text{control template})$

**Rank sum** score:
- Rank risk scores of all samples
- Sum up the ranks of the true PDs
Checking robustness of the risk marker

• Put one sample aside
  - Perform all the FS procedure on the remaining set - get a number n*
• Plot the distribution of n* obtained over all samples
• Note: this requires two nested LOOCVs
• Results: 4-20 features range; Majority of folds gave 8-9 features. 80% of the folds contained the full PD risk marker
LOOCV optimization of classifier on training set
Results

- Determined 8 as the optimal number of genes
- Used the training samples to determine the 8-gene signature
- Applied to the training samples - independent test samples
Molecular marker associated with PD risk

Scherzer C R et al. PNAS 2007;104:955-960

©2007 by National Academy of Sciences
Odds ratio

- Event occurs in a group with prob p
- **Odds** of the event: \( p/(1-p) \)
- **Odds ratio** – ratio of the odds in two groups

\[
\frac{p_1/(1 - p_1)}{p_2/(1 - p_2)}
\]

- Event: e.g. true PD
- **Groups**: top and bottom halves, or top/bottom %-iles
- **Cannot handle** events with prob 1
Results

• “The nominal odds ratio for PD of subjects in the third tertile was 210 [95% confidence interval (C.I.) 18–2,500] and in the second tertile was 18 (95% C.I. 2.0–150).”

• “The cross-validated odds ratios for PD were 5.7 (95% C.I. 1.6–21) and 2.2 (95% C.I. 0.6–7.8) for persons with scores in the third and second tertile, respectively.”

→Note the dramatic effect of not separating training and test set in the nominal OR!

→AUC values are shown but not reported - about 0.75
Results on an independent set

- 39 samples (19 PD, 5H, 15 DC)
- “There was a significant difference of scores in patients with PD versus healthy and disease controls ($P < 0.047$ by Wilcoxon test). High scores were significantly associated with increased PD risk ($P$ for trend $0.04$).
- Individuals with scores in the third tertile (high score) had an odds ratio for PD of 5.1 (95% C.I. 1-27), and individuals with a score in the second tertile (intermediate score) had an odds ratio of 1.9 (95% C.I. 0.4-9.6).”
Summary

• “Overall, the risk marker predicted PD beyond the prediction afforded by the traditional risk factors of age and sex ($P \leq 0.0001$ by Wald 2 test) and was not biased by dopamine replacement therapy”

• Can markers in the blood really identify disease in the brain? The jury is still out
Analysis of blood-based gene expression in idiopathic Parkinson disease

Ron Shamir, Christine Klein, David Amar, Eva-Juliane Vollstedt, Michael Bonin, Marija Usenovic, Yvette C. Wong, Ales Maver, Sven Poths, Hershel Safer, Jean-Christophe Corvol, Suzanne Lesage, Ofer Lavi, Günther Deuschl, Gregor Kuhlenbaeumer, Heike Pawlack, Igor Ulitsky, Meike Kasten, Olaf Riess, Alexis Brice, Borut Peterlin and Dimitri Krainc

First published September 15, 2017, DOI: https://doi.org/10.1212/WNL.0000000000004516
Idiopathic PD

• ~10% of the cases are genetic – known mutations can be tested
• >80% of the cases: idiopathic PD (IPD)
• Are IPD blood expression profiles distinguishable from healthy individuals?
Collected data

- After QC:
  - 205 IPD patients
  - 233 controls
  - 48 with other neurodegenerative disease (NDD)
- Measured whole blood expression profiles using Affymetrix arrays
**Analysis phases**

**Train:** compare analysis flows, classification algorithms
- Remove low intensity probes
- Remove samples from small batches
- Correct batch effects using fSVA
- Select N most differential probes
- Learn SVM classifier

**Validate:** evaluate models, tune parameters, select signature size. Determine final signature S

**Test:** final evaluation of S on new independent data

**Datasets used**

DS1. Training
- IPD: 140
- Controls: 153

DS2. Validation
- IPD: 35
- Controls: 40

DS3. Test
- IPD: 30
- Controls: 40
Demonstrating batch effects: Experimental setting.

An aside: batch effects

BL: baseline
FU: follow-up

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0156594
Red: RNA extracted and measured at BL,
Blue: RNA extracted at BL and measured at FU
dark grey: RNA extracted and measured at FU.
Batch effects in GenePark

- RNA collected in several clinical centers (Germany, France) over several years
- Periodically RNA extraction was done on groups of samples
- Those were sent to a central hub for processing and scanning
- Total of 40+ batches (location, time) of very variable size
- Understanding and correcting the batch effects was essential for getting results
- Used fSVA for batch correction
- Used leave-batch out cross validation
Leave-batch-out validation process

• The expression data show strong batch effects.
• The idea: validate the way we analyze the data by testing its ability to classify new samples from an independent batch
• The process:
  - Repeat:
    1. Exclude a complete batch
    2. Run the training process on the remaining samples
    3. Predict the labels of the excluded samples
  - Report the overall ROC score and accuracy
Leave-batch-out: results of the training process on DS1

Performance of the training process as a function of the number of probes selected for the signature

Good results for \( N = 30 - 100 \)

\( AUC = 0.72 \) for \( N = 50 \)
The next steps

- Train on the whole DS1 with N=20-200
- Test on DS2
- 100 probes gave the best results (AUC=0.79)
- Selected 100 probes as the final signature size
- Retrained SVM with 100 features on DS1+DS2 - obtained final signature S
- Applied S to DS3 (the test set)
ROC curves of the validation and test sets

Validation ROC: 0.79; Test ROC: 0.74
Prediction separately on old and new batches in the Test set

Old batches (46 samples) ROC: 0.765
New batches (24 samples) ROC: 0.789

New batches: unseen in DS1+2
Prediction on other NDD

Good separation
IPD-control, NDD-control.

Poor separation
IPD-NDD
Functional analysis of the signature genes.

Response to topologically incorrect protein (q = 0.033)

Cellular respiration (q = 0.015)

Polyubiquitination (q = 0.033)
## Comparing signatures from different studies

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Current study</strong></td>
<td>87</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Scherzer 2007</strong></td>
<td>0</td>
<td>22</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Soreq 2008</strong></td>
<td>0</td>
<td>0</td>
<td>53</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Grunblatt 2010</strong></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>136</td>
<td>3*</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Molochnikov 2012</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3*</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Shehadeh 2013</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>58</td>
<td>0</td>
</tr>
<tr>
<td><strong>Chikina 2015</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
</tr>
</tbody>
</table>

* Significant (p = 2.48E-5)
Summary

• Reasonably good IPD-control separation
• There is signal in the blood
• Robustness of the signature is an issue (as was shown in cancer)
• The holy grail: predict disease progression